1971-06

Investigation of speaker identification based on nasal phonation.

Young, Robert Bryant.

Monterey, California. Naval Postgraduate School

INVESTIGATION OF SPEAKER IDENTIFICATION BASED ON NAVAL PHONATION

Robert Bryant Young

United States
Naval Postgraduate School

THESIS

INVESTIGATION OF SPEAKER IDENTIFICATION
BASED ON NASAL PHONATION

by

Robert Bryant Young

Thesis Advisor: J. D. Campbell

June 1971

Investigation of Speaker Identification Based on Nasal Phonation

by

Robert Bryant Young
Lieutenant Commander, United States Navy
B.S., Boston College, 1958

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

from the

NAVAL POSTGRADUATE SCHOOL
June 1971

ABSTRACT

This thesis investigates the possibility of Speaker Identification through the use of Nasal Phonation. Short segments of a restricted set of words from one speaker were sampled, processed, and the resulting vector is used to represent the speaker. Representative vectors were formed for several speakers and correlated with vectors representing individual words from "test" speakers. The magnitude of the correlations of the word vectors with various speaker vectors were used to identify the speaker. This work expands on earlier work done in this field to the extent that it attempts to remove the subjective preparation of data and replace this instead with an objective process of computer mechanization. Some limited success was achieved and, just as important, critical problem areas are noted which, if improved upon as recommended, promise an improved identification capability. Two different word lists fundamental to the identification process were also investigated. Some data was obtained but it was not sufficient to suggest that one word list would be more productive than the other when used as the basis for speaker identification.

Recommendation is made to pursue further research in Speaker Identification using computer programming established during work on this thesis.

TABLE OF CONTENTS

I. INTRODUCTION ---------- 7

II. NATURE OF THE PROBLEM ---------- 8

A. SPEECH MECHANISM ---------- 8

1. Theoretical Basis of Experiments ------- 10

III. EXPERIMENTAL PROCEDURE ---------- 13

A. WORD LIST SELECTION AND RECORDING ---------- 13

B. DIGITAL CONVERSION ---------- 14

C. TAPE CONVERSION ---------- 16

D. POST CONVERSION PROCESSING ---------- 17

IV. EXPERIMENTS ---------- 20

A. EXPERIMENTAL RESULTS ---------- 23

V. CONCLUSIONS AND RECOMMENDATIONS ---------- 34

LIST OF REFERENCES ---------- 39

INITIAL DISTRIBUTION LIST ---------- 40

FORM DD 1473 ---------- 41

LIST OF TABLES

Table

I. Word Lists ---------- 21

II. Results of Experiment One ---------- 24

III. Results of Experiment Two ---------- 26

IV. Results of Experiment Three ---------- 27

V. Results of Experiment Four ---------- 29

VI. Results of Experiment Five ---------- 30

VII. Results of Experiment Six ---------- 31

VIII. Results of Experiment Seven ---------- 33

LIST OF FIGURES

Figure

1. Schematic Diagram of the Basic Speech Process -- 9

2. Sonogram of the Word "Nominal" ---------- 12

ACKNOWLEDGEMENTS

The author wishes to especially thank his Thesis Advisor, Assistant Professor J. D. Campbell, for his thoughtful guidance during the experiments and his assistance in the formulation of the thesis draft. Special thanks is also given to Assistant Professor V. M. Powers whose interest and enthusiasm for the thesis were warmly received.

I. INTRODUCTION

Much of what has been done in speech research in the past has resulted directly or indirectly from the dream of Homer Dudley back in the 1920's. It is recalled that Dudley wanted to transmit speech over a then newly-laid trans-Atlantic telegraph cable that had a bandwidth of little more than 100 cycles. In the process of trying to decode and reduce speech to its basic elements for narrow bandwidth transmission, much has been learned about the speech process itself and many useful and sometimes unexpected applications have appeared. One of these applications is in the field of speaker identification. Speaker identification, as the name implies, is an ability to recognize an individual speaker solely on the basis of his speaking traits. This goal has by no means been realized, chiefly because of the difficulty in isolating those speech parameters which are unique to a particular speaker.

In this thesis, an attempt has been made to pursue a very promising effort [Ref. 1] toward this goal which is based on Nasal Phonation. The reason for this approach can be gained by a brief examination of the human speech mechanism as we know it today.

II. NATURE OF THE PROBLEM

A. SPEECH MECHANISM

It has been established through research by Flanagan, *et. al.*, [Ref. 2] and others that sound can be generated in the vocal system in three ways. Voice sounds are produced by elevating the air pressure in the lungs, forcing a flow through the vocal cord orifice (the glottis) and causing the cords to vibrate. The interrupted flow produces quasi-periodic, broad-spectrum pulses, which excite the vocal tract. Fricative sounds are generated by forming a constriction at some point in the tract, usually toward the mouth end, and forcing air through the constriction at a sufficiently high Reynolds number to produce turbulence. A noise source of sound pressure is thereby created. Plosive sounds result from making a complete closure, again usually toward the front of the mouth, building up pressure behind the closure and abruptly releasing it. All these sources are relatively broad in spectrum. The vocal system acts as a time-varying filter to impose its spectral characteristics on the sources.

With reference to anatomy, the major parts of a man's vocal apparatus are shown schematically in Figure 1. The vocal tract proper is a non-uniform acoustic tube about 17 cm. in length. It is terminated at one end by the vocal cords (or by the opening between them, the glottis) and at the other end by the lips. The cross-sectional area of the

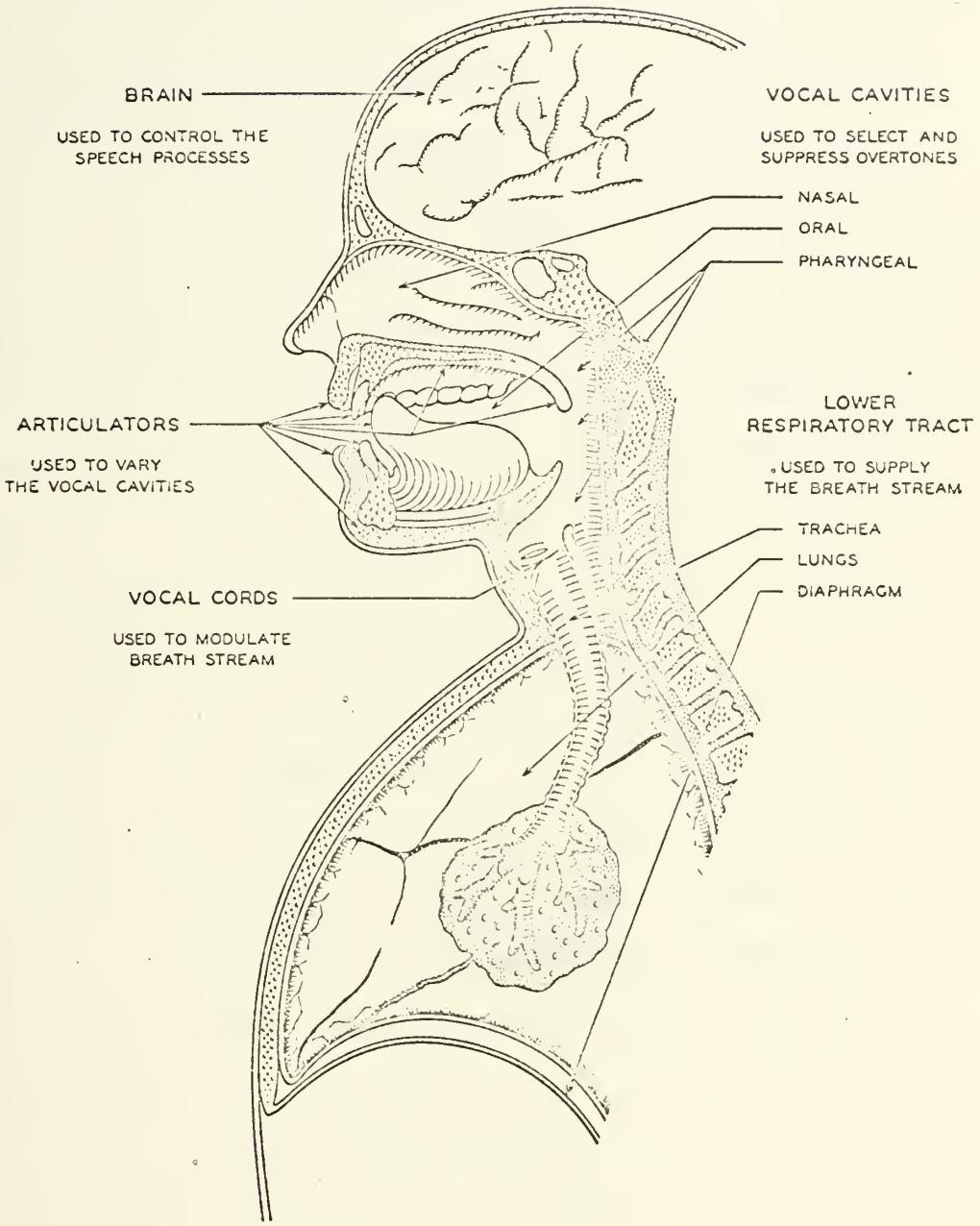

Figure 1. Schematic Diagram of the Basic Speech Process.

tract is determined by placement of the lips, jaws, tongue, and velum and can vary from zero (complete closure) to about 20 $cm.^2$.

An ancillary cavity, the nasal tract, can be coupled to the vocal tract by the trapdoor action of the velum. The nasal tract begins at the velum and terminates at the nostrils. In man, this cavity is about 12 cm. long and has a volume of about 60 $cm.^3$. In non-nasal sound, the velum seals off the nasal cavity, and no sound is radiated from the nostrils. It is the sound produced when this cavity is coupled to the vocal tract that is of particular interest in speaker identification based on nasal phonation.

1. Theoretical Basis of Experiments

In the search for the parameters which might differentiate a given speaker from another, one is motivated by the desire to find a simple and compact solution which will ultimately lend itself to practical implementation. Therefore, one is prompted to investigate parameters which would be consistently available from only a very small sample of speech. A small sample can be a short-time duration. In this respect, nasal phonation is an attractive approach because it is based on data derived from nasal consonants which occur in certain words for short time periods of from 35-50 milli-seconds depending on the speaker. Analysis has shown that during their production acoustic radiation occurs through the nostrils with the closed oral cavity acting as a shunt. The articulators do not move

during the period of closure, and the vocal cavities remain nearly fixed. Hence, the power spectrum of radiated acoustic energy is essentially steady as indicated by the sound spectrogram of the word "nominal" in Figure 2. Note that the formants show very little movement for the duration of the two nasals. Researchers have commented on the marked extent to which the acoustic properties of nasal consonants vary from speaker to speaker [References 3 and 4]. If these indications are true over a wide population of speakers, it is possible, then, that acoustic radiation produced during the phonation of nasal consonants could provide a strong clue to speaker identity.

Another important feature of nasal consonants with respect to speaker identification is the relative frequency with which they are used in spoken English. According to Tobias [Ref. 5], nasals comprise 11% of the phoneme content of commonly spoken English. Thus, there is sufficient data available in a short sample of speech to provide some degree of speaker identification if our hypothesis is correct. In substantiation of this approach, Glenn and Kleiner [Ref. 1] have reported that in experiments involving a population of ten speakers an average identification accuracy of 97% was obtained. With an experimental population of thirty speakers, identification accuracy was 93%.

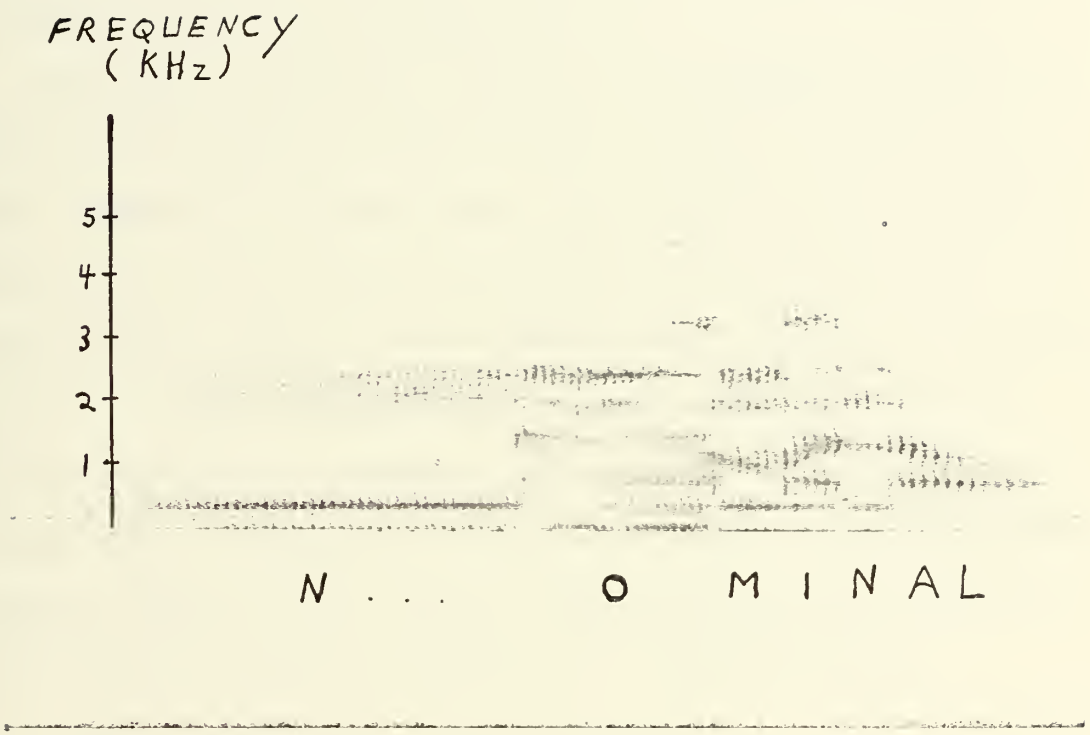

Figure 2. Sonogram of the Word "Nominal".

III. EXPERIMENTAL PROCEDURE

The approach taken in this work is along the lines of Glenn and Kleiner. However, an attempt has been made to use the computer to a greater extent for the processing and manipulation of data and to develop facets of the problem not covered in their experiments.

The basic analysis technique used throughout the experiments described in this thesis is the reduction of a time segment (the near steady state portion during nasal phonation) of selected words into normalized spectral components which form a vector describing a speaker. The same process is again performed on another different set of words spoken by an "unknown" speaker; another vector is formed and is compared with known speaker vectors using the highest value of the cosine of the angle between the compared vectors as the criterion for identification.

A. WORD LIST SELECTION AND RECORDING

Various word lists were used in the different experiments and each will be specifically discussed under the heading of the applicable experiment. However, in each case a word list contained twenty different words beginning with the nasal consonant "n". A given speaker was usually required to record twenty words in slow sequence. The microphone used for this recording was a SURE MODEL 575S. The dynamic range was sufficient for the 1 - 3.5 KHz area of spectrum interest. The recorder used was an Ampex SP-300,

the only one conveniently available. As was later indicated, this recorder caused some problems because of the introduction of high background noise. Since noise bursts could not be tolerated on the tape (because of sensitivity of logic sensing circuits in the digitizing process), considerable care had to be taken during recording. In addition to a rehearsal of the word list, each speaker was cautioned to avoid movements of the microphone, tapping, and other motions, which might introduce such bursts. Also, as it was only possible to use a single recorder output, a procedure was devised whereby the VU meter could be used to judge word placement on the tape during the digitizing process. This procedure required the speaker to hum a tone before and after the word list. This tone could be easily distinguished on the VU meter and greatly facilitated the digitizing process. Several seconds of silence were allowed between words and after the initial tone.

B. DIGITAL CONVERSION

The recorded words were then passed through a 1 - 3.5 KHz Khron-Hite Series 3320 filter and then input to a CI 5000 Analog Computer. Here, after amplification, logic sensing circuits determined the start time of the digitizing process which was performed at 12.5 KHz by an SDS 9300 digital computer.

The start time of the digitizing process is of critical concern because of the need to sample the spoken nasal

consonant during the proper interval -- i.e., when the radiated acoustic energy is most nearly in steady state. As expected, there is a small transient period initially, then near steady state. Fujimuras' analysis [Ref. 3] has shown varying degrees of formant variation toward the tail end of steady state depending on the following vowel. Thus, there appears to be an optimum steady-state window extending for approximately ten milli-seconds during the time of consonant. The logic sensing arrangement allowed for a variable delay to preceed the start of the digitizing process after triggering on the start of a word. Additionally, there was provision of a delay flop to insure that once triggered, the sensing circuit would be immune from further (false) triggers for a predetermined amount of time (experience has shown this time should be about 1.5 seconds).

The number of samples taken was fixed for this processing arrangement at 128 which, at a sampling rate of 12.5 KHz, takes about ten milli-seconds. A ten milli-second delay after triggering was used to insure that the sampling period lay within the steady state optimum sampling window. The choice of sampling frequency also allows a convenient feature (as will be shown later) of having the Fourier coefficients centered in approximately 100 cycle bandwidths.

Thus, the digital output of the conversion for one word list by one speaker consists of twenty blocks of 128 samples each, which are recorded on a seven-track digital

tape as a single file of data. In actual practice, the process was repeated twice again for each file to insure that each word was properly sampled at least once.

In early tests, files were often discarded because it was found that they were not complete. A typical reason for these errors was noise bursts on the voice tape. Though attempts were made to keep this type of error to a minimum, there were still circumstances which required manipulation around these bursts, if data was to be saved. If the burst was distinguishable from the spoken words and occurred prior to the twenty words, it was possible to delay energizing the logic recognition/delay circuits until immediately prior to the first spoken word. The timing was physically difficult to implement in many instances and hence, the requirement to make several runs of the same data to ensure success.

C. TAPE CONVERSION

Since the SDS 9300 did not have the digital storage needed for analysis, use was made of the school's IBM 360. Because the current operating system supports only nine-track tape for FORTRAN input files, it was necessary to convert from the seven-track tape (produced in the digitizing process) to a nine-track tape. To facilitate evaluation of data, a decimal print out of data by block was also provided. Thus, for a given experiment, the entire content of the seven-track tape was converted (included

redundant files); the decimal data was evaluated for completeness, and then, through the use of Job Control Language (JCL), selected files were brought forward and converted for further processing. It should be noted that this editing process allowed mainly a quantitative check on the data; it did not readily permit comparison of this data with the original analog data.

D. POST CONVERSION PROCESSING

Because the energy distribution by frequency was required, the next step in the process was to determine the spectral content of the data blocks. The Fast Fourier Transform Algorithm [Ref. 6] was used to compute the complex Fourier coefficients whose magnitude squared are the energy spectrum. Since a 12.5 KHz sampling rate was used, each value of the energy spectrum represented an incremental bandwidth of 97.66 Hz (very close to the 100 Hz bandwidths that were manually quantized by Glenn and Kleiner). Again, closely paralleling Glenn and Kleiner's work, an approximately 2.5 KHz band of the spectrum was isolated by discarding those values of the energy spectrum below 1025.36 Hz and above 3466.83 Hz. This spectrum band of interest then included twenty-five segments of 97.66 Hz, each of which was considered a component of a twenty-five dimensional vector which represented a particular sampled word. Each of the twenty-word vectors then underwent a normalization transformation according to the formula:

$$v_i' = v_i \Big/ \sum_{j=1}^{25} v_j \ , \quad i = 1,2,\ldots,25.$$

Additionally another transformation was performed on each vector which was designed to emphasize the major pole and the major zero of the power spectrum. For the vector

$$V' = (v_1', v_2', \ldots, v_{25}').$$

Let

$$M = \max \{v_i'\}$$

represent the major pole of the spectrum, and

$$m = \min \{v_i'\}$$

the major zero. Then the transformed vector is given by

$$V^* = (v_1^*, v_2^*, \ldots, v_{25}^*)$$

where

$$v_i^* = \alpha v_i' - \beta$$

with

$$\alpha = 1/(M-m)$$

$$\beta = m/(M-m).$$

The vectors thus transformed will be called subvectors. At this point, for the purpose of the experiments, the first ten word subvectors were considered to have derived from a known speaker and the second ten word subvectors to have been uttered by an unknown or "test" speaker. Each set of ten subvectors was then arithmetically averaged by component

forming two "prime" vectors (representing reference and test vectors for each speaker). For a given experiment involving several different speakers, the cosine of each test vector with each reference vector was computed. The highest value in this correlation process was considered to have identified the test vector with the speaker. If the test vector with highest correlation was in fact from the same speaker, the result was judged a "match."

Throughout the data manipulation process, program output/input was returned to magnetic tape. However, at the point where subvectors were generated, punched cards were used to retain the data. The use of cards allowed flexibility in further analysis of data.

This analysis took the form of correlation between each subvector and the prime vector thereby generated. Also, differences were taken between components of subvectors and prime vectors so that frequency ranges exhibiting the greatest component difference could be isolated. Finally, additional programming was established to "edit" prime vectors by reducing the number of subvectors generating them. Of course, facility was retained to perform correlations and component differences as was done with the case of the prime speaker vector being generated from a full ten subvectors.

IV. EXPERIMENTS

With the computer programming mechanization established, several experiments were undertaken to explore some variations in speaker identification problem parameters.

Because of Fujimuras' work [Ref. 3], which has shown varying degrees of formant variability toward the end of the near steady state condition depending on the following vowel, it was decided to examine lists of words categorized by a particular vowel following the nasal consonant "n". The nasal consonant was chosen at the beginning of the word in order to facilitate digitization, thereby avoiding the problems of detection of the nasal consonant within the word itself. Initially, three word lists were chosen for the vowel sounds i as in beet, ε as in set and æ as in sat. Two additional word lists were added to gain further information. These were a list containing a mixture of words from lists one through three and finally a list containing words with various randomly selected vowels following the nasal consonant. These five lists are shown in Table I. It will be noted that many of the words listed are not meaningful words in the English language. These words were fabricated solely for the purpose of filling each list to twenty words. A single speaker (the author) was used in the initial experiments.

The goal of Experiment One was to decide on a "best" word list for later identification studies with a number of

TABLE I.

Word Lists

	1	2	3	4	5
1	neat	net	nag	nee	nice
2	neeg	neb	nanny	Negro	nimble
3	neap	nest	nap	neon	Nyquist
4	near	nectar	narrate	neef	nominal
5	neek	negative	napkin	neem	numbering
6	nee	nekton	narrow	neev	nonsense
7	need	nelson	nastily	net	normal
8	needle	nemesis	nasty	nest	nation
9	needless	nephew	natty	nectar	Nixon
10	Negro	nepotism	natural	Nelson	notion
11	neither	nebula	Navaho	ness	nobile
12	neolithic	Neptune	navigate	next	new
13	neology	ness	Nazarene	nag	nanny
14	neon	neck	Nazi	nanny	notice
15	neophyte	nestle	nano	nap	north
16	Negress	nether	naphtol	narrate	nine
17	Neartic	network	naphtha	napkin	numeral
18	neef	never	nat	narrow	national
19	neem	next	national	nasty	noel
20	neev	nef	naf	never	notify

different speakers. As a criterion for judgement, the second set of ten words in each list was considered a "test" word list. The "best" word list was to be chosen on the basis of maximum correlation in terms of cosine values between the "reference" and "test" word lists.

In Experiment Two words from the first word list were spoken by five different male speakers. The vectors obtained from these words were used to form reference and test prime vectors which were then compared in the correlation process described previously.

Experiment Three repeated Experiment Two except that the logic circuit delay from the start of each word was set at one milli-second vice ten milli-seconds.

Experiment Four used all the basic data of Experiment Two except that the ten subvectors used to form each prime vector for both the Reference and Test speaker, were "edited" to seven subvectors. The three subvectors removed in each case were those whose correlation with the original prime vector were smallest in value. Those vectors removed ranged in this correlation between .4443 and .7666. It was noted that this upper value along with two other .7... correlation values (of six total) were removed from speaker three's Reference and Test vectors. The inconsistent removal of subvectors with large correlation values tends to degrade the data. This is further discussed under Conslusions and Recommendations.

Experiment Five used all the basic data of Experiment Three except that again the three lowest correlating

subvectors with each prime vector were "edited." These subvectors removed ranged in correlation between .2564 and .7814.

Experiment Six was an attempt to identify thirteen male speakers.

Experiment Seven used the same basic data of Experiment Six except that the lowest three correlating subvectors were edited from each prime vector. Three subvectors removed ranged in correlation between .2625 to .8639.

A. EXPERIMENTAL RESULTS

Table II lists the results of Experiment One. As can be seen from the Table, the match between reference and test vectors for the first word list is high and possesses significant emergence. Emergence is here defined to mean the relatively large value of correlation of a test vector with a given reference vector as compared with other reference vectors. This emergence quality was not obvious in other word lists and for this reason, it was decided to use this list in later identification experiments.

Word list three was also interesting. Firstly, the extreme high correlation between word list three and test word list four should be noted (this value was in fact the highest recorded in the table). The fourth word list, which it will be recalled was a mixture of word lists one through three, was not mixed well because, of the last ten words (of test list four), eight were repeated from Reference

TABLE II

Results of Experiment One

Reference Word List

Test Word List	1	2	3	4	5
1	.9137	.3537	.3245	.6496	.3199
2	.3847	.8849	.9432	.7507	.8303
3	.4503	.8885	.9633	.8163	.7960
4	.3978	.8989	.9661	.7631	.6937
5	.3491	.7185	.8504	.6865	.8603

List three. This probably explains the high correlation value with that list. If, then, this value is excepted, word list three correlated with test word list three extremely well; in fact, the best of those tested (although it does not have the nice emergence property found with word list one). Because of this high correlation, it was decided to use word list three in Experiments Six and Seven.

The results of Experiment Two are shown in Table III. As can be seen, clear matches are obtained for speakers two and four. Speakers one and five were near matches, the correct match in each case being the second highest correlation value. Speaker three is a poor match placing third in correlation but with a significant magnitude difference from the highest correlation. In the search to improve the number of matches, a review of the experiment diary disclosed that speaker two (the author) and speaker four had both spoken very slowly.

It was thought that perhaps because of the speed of the speech the steady state window was missed. Accordingly, the experiment was rerun using a shorter logic delay (one milli-second).

The results of Experiment Three are given in Table IV. This time clear matches were obtained for speakers two, four, and five. Speaker one again placed second for highest correlation and speaker three showed poorest placing out of the five speakers. The reason for speaker three's

TABLE III.

Results of Experiment Two

Test Speaker	Reference Speaker 1	2	3	4	5
1	.9341	.1886	.6185	.2989	.4954
2	.3221	.9625	.7365	.7100	.7594
3	.9357	.2880	.7397	.5044	.6243
4	.4203	.6633	.8499	.9781	.9257
5	.8555	.4212	.8873	.7296	.8583

TABLE IV.

Results of Experiment Three

Test Speaker	Reference Speaker 1	2	3	4	5
1	.9164	.1828	.6998	.4863	.7683
2	.3750	.8209	.7028	.6188	.6766
3	.8243	.2121	.6324	.3497	.6580
4	.5117	.7055	.8430	.9299	.8467
5	.9196	.4826	.8894	.7988	.9514

poor correlation is unknown. It was noted in the diary that the speaker's words were close together, and it is possible that the logic circuit did not have sufficient opportunity to stabilize prior to a following word thus introducing false data into the experiment.

The results of Experiment Four are given in Table V. As can be seen, the editing process did improve the match characteristics of the data set. However, the correlation value of speaker five with reference five has now been reduced to a poor third. Speaker three's correlation values have not measurably improved. This might be expected because of the removal of only high correlation subvectors in the editing process.

The results of Experiment Five are shown in Table VI. Again the overall match record is three out of five. However, speaker one was now matched and speaker four unmatched, although it should be noted that speaker four places second. Speaker three is still the least likely match as was the case in the unedited version of the experiment (Experiment Three).

The results of Experiment Six are given in Table VII. There are six clear matches. Of the remaining seven, two miss matches by one, i.e., place second, and one (speaker twelve) placed third. The worst case was the speaker who was mismatched by seven. It is interesting to note that the recording diary had the comment "a little fast" for this speaker's recitation. This means that the speaker

TABLE V.

Results of Experiment Four

Test Speaker	Reference Speaker 1	2	3	4	5
1	.9176	.1175	.4491	.1604	.2906
2	.1854	.9483	.7105	.6391	.7679
3	.9095	.2628	.7076	.4981	.5087
4	.2586	.5276	.8399	.9369	.9526
5	.8627	.3447	.8073	.7110	.7136

TABLE VI.

Results of Experiment Five

Test Speaker	Reference Speaker 1	2	3	4	5
1	.9064	.1647	.6169	.5735	.7994
2	.2556	.6771	.7044	.4809	.6503
3	.6148	.1361	.3950	.2903	.6096
4	.3411	.6397	.8069	.7423	.7381
5	.8691	.5194	.8263	.8346	.9542

TABLE VII.

Results of Experiment Six

Test Speaker	Reference Speaker 1	2	3	4	5	6	7	8	9	10	11	12	13
1	.9586	.7587	.4738	.8944	.3481	.9086	.8742	.8025	.9730	.6485	.9778	.9884	.4585
2	.6063	.9818	.3968	.8277	.8041	.8545	.6461	.5103	.7205	.9284	.7535	.7670	.9383
3	.6704	.5680	.9053	.6954	.3695	.6384	.7603	.7510	.6672	.4002	.6491	.6166	.3667
4	.9471	.6773	.5434	.8889	.4166	.7965	.8191	.9611	.9158	.4664	.9085	.8308	.3999
5	.4791	.8699	.3029	.7635	.7774	.7491	.4630	.4571	.5574	.9000	.5915	.5606	.9494
6	.9578	.6505	.4590	.8229	.2760	.8176	.8531	.8421	.9477	.4872	.9542	.9292	.3111
7	.9246	.7923	.5414	.9023	.4891	.8765	.9179	.8267	.9300	.6469	.9498	.9419	.5202
8	.8811	.5202	.6098	.7925	.2852	.6860	.7202	.9659	.8025	.3385	.7857	.6871	.2794
9	.9847	.6045	.5192	.8667	.2672	.7974	.7983	.9311	.9148	.4432	.9045	.8561	.3107
10	.6739	.9398	.3251	.8713	.6070	.9384	.6266	.5584	.7898	.9762	.8067	.8108	.8840
11	.9431	.7965	.5457	.9053	.4022	.9270	.8682	.8576	.9870	.6615	.9818	.9566	.5115
12	.9570	.6666	.4851	.8339	.2939	.8284	.8657	.8294	.9525	.5138	.9585	.9430	.3343
13	.6882	.9253	.3826	.8903	.5712	.9356	.6778	.5566	.7850	.9445	.8024	.8434	.8376

said the words quickly, not emphasizing the nasal consonant. Further, this was the only speaker in this experiment for which such a comment was made.

The results of Experiment Seven are given in Table VIII. There is a sharp reduction in matches as a result of the editing process; only three matches in thirteen were obtained.

Two of the three matches were also obtained in Experiment Six; the other match (speaker four) was a fifth place match in Experiment Six.

The absence of the hoped-for improvement in number of matches as a result of the editing process was disappointing; however, it is understandable in view of the imbalance in selection of the subvectors rejected during editing.

This will be discussed further in the following section.

TABLE VIII.

Results of Experiment Seven

Test Speaker	Reference Speaker 1	2	3	4	5	6	7	8	9	10	11	12	13
1	.9707	.7277	.8848	.7374	.1868	.8447	.9009	.6310	.9508	.4698	.9637	.9499	.3479
2	.5521	.9444	.2580	.8063	.7531	.7930	.5770	.4182	.5801	.8842	.6809	.7515	.9336
3	.4019	.4068	.8545	.5384	.3259	.4763	.5920	.5897	.4454	.2729	.4368	.4555	.3030
4	.9298	.6597	.1870	.8469	.3537	.7856	.8161	.8982	.9351	.3983	.8956	.8402	.3857
5	.3128	.8154	.1773	.7104	.7265	.6422	.3420	.2818	.3244	.9551	.4880	.5102	.9746
6	.9830	.6335	.1054	.6981	.1677	.7643	.8977	.6728	.9397	.3593	.9553	.9027	.2530
7	.9310	.7389	.2402	.7713	.3254	.8199	.9491	.6347	.8865	.5332	.9559	.9227	.4379
8	.7978	.4533	.2596	.7531	.2143	.6282	.6355	.9694	.7969	.2491	.6803	.6376	.2364
9	.9706	.5872	.1557	.7825	.1764	.7479	.8421	.8250	.9388	.3238	.8813	.8517	.2465
10	.5113	.9561	.1267	.8131	.4589	.8852	.4615	.3705	.5795	.9446	.6079	.7619	.8767
11	.9440	.7194	.2015	.7742	.2491	.8562	.8522	.7799	.9630	.4747	.9588	.8990	.3891
12	.9727	.6197	.1033	.6765	.1712	.7427	.9058	.6348	.9188	.3649	.9570	.8912	.2528
13	.5607	.9632	.2045	.8220	.4729	.8954	.5777	.3651	.6242	.8814	.6436	.8263	.8285

V. CONCLUSIONS AND RECOMMENDATIONS

It is considered that the results of the experiments show there is some merit to the use of nasal phonation for speaker identification. Although the number of speakers was limited and the absolute match percentage did hover at about 50%; nevertheless, it is believed that there may be instances of closed groups of speakers where the ability to obtain a 50% credible match would be a useful adjunct to other information toward identifying that speaker. Looking at the 50% correct identification from another point of view, it should be noted that, for example, in Experiment Six, the data also showed that for a given match, the correlations obtained would be useful in eliminating about one third of the speakers from consideration in the identification process.

It must be remembered that the process used in these experiments for the actual extraction of the subvectors was done strictly by machine whereas Glenn and Kleiner [Ref. 1] used manual methods. It was the intent of this investigation to remove all subjectivity from this process because of having experienced considerable error from previous work with subjective sonogram comparisons.

The computer mechanization, as now established to handle this analog voice data, is sound; and, it is believed, that with further refinement, many inaccuracies now inherent in the process can be removed; thereby yielding the promise

of a higher percentage of correct identifications, even with a larger population of unidentified speakers.

The most critical point in the digitizing process is the method used to start the digitizing. As was pointed out previously, the start time is sensitive to noise and one must have a check to ensure that it is done properly. Under the present scheme, reliance was placed on a delay flop to start the process. However, what was not taken into consideration was the problem of noise in the tape recorder. It is probable that because of this noise threshold there were instances when the analog voice voltage was beneath the noise for a fraction of time prior to delay timing by the flip-flop circuit. The particular recorder used for the experiments (as mentioned earlier) was noisy and hence, inconsistencies in start of sampling may have resulted.

One way to improve this situation is to use a more noise-free recorder not available for this project. Also, there should be a simultaneous graphical record of the analog voltage from the recorder along with the digital to analog record of the digitized version. For convenience, a milli-second time tick could also be plotted. This would assist the operator on a near real time basis to ensure that data is taken at the desired time. Yet another improvement would be the use of more sophisticated logic to start the digitizing. Such logic might use some type of nasal consonant recognizer to insure that digitizing starts during the steady state.

A second area needing improvement is the editing process. Even with the improvement for start of digitization, there is still the possibility that extraneous recorded sound energy not related to the desired signal will be present in the subvectors. Hence, the need to edit them. In the experiments as described, the first criterion for a "bad" subvector was its poor correlation with the prime vector it helped to generate. These subvectors were discarded mainly on this basis. However, upon closer analysis of the twenty-five component differences between the subvectors and prime vectors, it was found that, although poor correlations could be relied upon in most cases as an indication for rejection, there were other cases where there were irregularities in components which compensated one another and hence, gave a higher correlation value and, therefore, were retained. These latter subvectors should also have been rejected for they were probably sampled for one reason or another during a non steady state condition.

Another area of error overlooked in the editing process was the decision of how many subvectors to eliminate. Three were arbitrarily chosen for each speaker mainly because it was convenient for computer processing and also it was felt that this small number would not drastically affect the true character of each prime vector. However, as was seen in the experiments, when this selection rule was applied to all speaker subvectors, some high correlating

subvectors are removed and hence the data is distorted. The discarding of subvectors should be made solely on the basis of a small magnitude correlation value and, as mentioned above, of component variance.

All speaker data was retained throughout the experiments, and except for the editing described above, no attempt was made to "dress" the results by discarding data which was often misclassified. Speaker three in experiments Two through Five was one that might have been so eliminated. This particular data was suspect because of his fast speech and consequent uncertainty regarding proper digitization. Speaker six in Experiments Six and Seven was another example.

Another interesting fact is that words spoken by speaker two (the author) were always correctly identified (at least prior to the editing process). Though the data base is slim, it suggests that greater care in recording of words spoken by individuals taking part in the experiment would yield measurably improved results.

The best word list to use is still open to question. Data produced from the experiments does not favor either of the two lists studied and in any case is not exhaustive enough for a firm decision.

Lastly, the author would like to point out that, although it was merely a tool to an end, the computer mechanization and debugging of the various programs used for the experiments was formidable. It is hoped that interest in Speaker Identification will continue at the Naval Postgraduate School and that the establishment of the programming

used in this thesis will be used as a stepping stone to further research.

LIST OF REFERENCES

1. Glenn, J. W. and Kleiner, N., "Speaker Identification Based on Nasal Phonation," The Journal of the Acoustical Society of America, v. 43, p. 368-372, February 1968.

2. Flanagan, J. L., Coker, C. H., Rabiner, L. R., Schafer, R. W., and Umeda, N., "Synthetic Voices for Computers," IEEE Spectrum, v. 7, p. 22-45, October 1970.

3. Fujimura, O., "Analysis of Nasal Consonants," The Journal of the Acoustical Society of America, v. 34, p. 1865-1875, December 1962.

4. Dickson, D. R., "An Acoustic Study of Nasality," Journal of Speech and Hearing Research, v. 5, p. 103-111, June 1962.

5. Tobias, V., "Relative Occurrence of Phonemes in American English," The Journal of the Acoustical Society of America, v. 31, p. 631, 1959.

6. Cooley, J. W. and Tukey, J. W., "An Algorithm for the Machine Calculations of Complex Fourier Series," Mathematics of Computations, v. 19, p. 297, April 1965.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Asst Professor J. D. Campbell, Code 52Cb Department of Electrical Engineering Naval Postgraduate School Monterey, California 93940	2
4.	LCDR Robert B. Young, USN USNAV SEC GRU ACTY FPO, New York 09513	5

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

INVESTIGATION OF SPEAKER IDENTIFICATION BASED ON NASAL PHONATION

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)
Master's Thesis; June 1971

5. AUTHOR(S) (First name, middle initial, last name)
Robert B. Young

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1971	42	6

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

This thesis investigates the possibility of Speaker Identification through the use of Nasal Phonation. Short segments of a restricted set of words from one speaker were sampled, processed, and the resulting vector is used to represent the speaker. Representative vectors were formed for several speakers and correlated with vectors representing individual words from "test" speakers. The magnitude of the correlations of the word vectors with various speaker vectors were used to identify the speaker. This work expands on earlier work done in this field to the extent that it attempts to remove the subjective preparation of data and replace this instead with an objective process of computer mechanization. Some limited success was achieved and, just as important, critical problem areas are noted which, if improved upon as recommended, promise an improved identification capability. Two different word lists fundamental to the identification process were also investigated. Some data was obtained but it was not sufficient to suggest that one word list would be more productive than the other when used as the basis for speaker identification.

Recommendation is made to pursue further research in Speaker Identification using computer programming established during work on this thesis.

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Nasal Phonation						
Speaker Identification						

Thesis 133050
Y69 Young
c.1 Investigation of
speaker identification
based on nasal phona-
tion.